INTEGRACIÓN DE UN CORPUS DE TEXTOS BILINGÜE Y UN GLOSARIO DEL CAMPO DE LA INFORMÁTICA

Mº DEL SOCORRO BERNARDOS GALINDO
GUADALUPE AGUADO DE CEA
Universidad Politécnica de Madrid

RESUMEN

En este trabajo se presentan las principales características de Calíope, una aplicación web que es capaz de manejar un corpus y un glosario de términos en inglés y en español. La singularidad más importante de esta herramienta es que permite interrelacionar estos dos recursos. Así, por ejemplo, los resultados de la búsqueda de concordancias se pueden incorporar automáticamente a los ejemplos de uso del término correspondiente en el glosario; y desde la lista de palabras de un texto del corpus se pueden añadir términos al glosario o acceder a la información de un término que esté en el glosario.

Palabras clave: lingüística de corpus, terminología informática, fraseología especializada

ABSTRACT

The aim of this paper is to present the main features of Caliope, a web application which is able to manage a corpus and a glossary of terms in English and Spanish. The most important singularity of this tool is that it supports the interrelation between these two resources. Thus, for instance, the results from the search of concordances can be incorporated into the examples of use of the corresponding term in the glossary; and a term from the list of words of a text in the corpus can be automatically added to the glossary or accessed if it is already there.

Keywords: corpus linguistics, computer terminology, specialized phraseology

1. INTRODUCCIÓN

Desde la incorporación de la lingüística de corpus al estudio de las lenguas muchas han sido las aplicaciones desarrolladas para fines tanto didácticos como investigadores (Sánchez 2000), que engloban aspectos tan variados como: creación de diccionarios (Sinclair 1987; de Schryver y de Pauw 2007), análisis de las características del inglés académico (Perales-Escudero y Swales 2011), análisis de errores de aprendizaje (Granger 2011), corpus multilingüe del IULA[1], entre otros campos. Sin embargo, encontrar aplicaciones que integren en una misma herramienta varias de estas finalidades, no es tarea fácil.

En nuestro caso, se pretende dar respuesta tanto a las necesidades terminológicas, como lingüísticas de alumnos universitarios del campo de la informática que necesitan mejorar tanto las destrezas de escritura académica y profesional en inglés como en español. Las metas eran integrar en una misma aplicación, Calíope, (Aguado y Bernardos 2007), tanto las prestaciones proporcionadas por un concordanciero convencional, que facilitara al alumno el aprendizaje de los términos en contexto y le permitiera ver las relaciones sintácticas y léxico-semánticas que se establecen entre ellos, como disponer de un diccionario de especialidad bilingüe y un corpus de textos de los diferentes sub-dominios de la informática en ambas lenguas, así como gestores computacionales para ambos recursos.

En la actualidad existen numerosos gestores de corpus textuales (por ejemplo, BwanaNet[2], View[33] y WordSmithTools[44]), de glosarios y bases de datos terminológicas (por ejemplo, DiCoInfo[55] y Termium Plus[66]), y, en menor medida, de ambos tipos de recursos (por ejemplo, Terminus 2.0[77]), cada uno con distintas características. Ahora bien, cabe destacar que en el momento en el que comenzó el proyecto de Calíope no había disponible ninguna aplicación de este último tipo, que integrara ambos recursos y permitiera establecer interacciones entre ellos, ni era posible integrar los gestores de cada recurso por separado del modo que se necesitaba.

En el resto del artículo se describe la situación actual de Calíope, se explica su última versión, la 3.0, y los aspectos en los que se está trabajando o se va a trabajar en un futuro.

Si se tiene en cuenta la integración de un modelo lingüístico en la concepción del sistema, merece destacarse DiCoInfo, *Le dictionnaire fondamental de l'informatique et de l'internet*, desarrollado por el equipo de Marie Claude L'Homme, siguiendo los principios de la Teoría Sentido-Texto (TST) (Melcuck et al. 1988). Este diccionario desarrollado inicialmente para francés e inglés (aunque ahora cuenta también con español, pero con menor número de términos) está destinado a cualquier usuario interesado por la informática. El hecho de estar asentado sobre los modelos lexicológicos de la TST, incorporando tanto la estructura actancial y los actantes como los papeles semánticos de la gramática funcional, y las funciones léxicas, lo convierte en una herramienta lexicográfica indispensable y muy completa para los seguidores de la TST, así como para los estudiosos de las diferentes corrientes en lexicografía especializada. No cabe duda de que DiCoInfo es un recurso imprescindible a la hora de realizar estudios terminológicos en el campo de la informática e Internet. Ahora bien, al modificar los rasgos pragmáticos previos a la confección de una herramienta de este tipo, como destinatarios, finalidad, conocimientos previos, necesidades de los usuarios, etc., necesariamente los objetivos que se persiguen han de ser diferentes.

2. DESCRIPCIÓN DE CALÍOPE

Calíope, como cualquier otra aplicación web, consta de un conjunto de páginas con las que se interactúa navegando a través de los menús principales que componen dichas páginas, y de una serie de elementos como botones específicos, pies de páginas, formularios, etc.

Las funciones disponibles serán distintas según los privilegios que tenga el usuario que esté utilizando la aplicación. Si se tienen permisos de administrador, éste podrá acceder a todas las páginas de la aplicación. En caso contrario, sólo podrá realizar las operaciones de consulta de *Calíope*.

Para acceder a *Caliope* se debe introducir nombre de *usuario* y *contraseña.* Si los datos introducidos son correctos, se desplegará la ventana principal de la aplicación con los siguientes menús: *administrar usuario, consultar corpus* y *consultar glosario*. En la parte superior se encuentran los iconos correspondientes a los idiomas

español e *inglés*, que permiten al usuario cambiar el idioma de la interfaz cuando desee.

Con la versión actual (Martín 2012) se pueden realizar tres tipos de consultas sobre el **corpus**: *visualización de un texto*, *listado de palabras*, *concordancias*/ *coocurrencias*. Para el **glosario** de términos se pueden *consultar* los *términos* y la *información* asociada a cada uno de ellos. Además, el administrador podrá gestionar ambos recursos, junto con los **usuarios** y el **histórico** del sistema. En las siguientes subsecciones se detallan cada una de las operaciones.

2.1 Operaciones sobre el corpus

Los textos que se quieren **visualizar** pueden filtrarse según los siguientes parámetros: idioma, título, campo, tipo, año de publicación, autor y usuario que lo introdujo. Además se pueden elegir los registros por página (número de textos que aparezcan en la pantalla principal) (Véase Figura 1). Para verlo basta con pinchar sobre el nombre del texto deseado. En la parte central aparecerá el texto seleccionado y en la parte de la izquierda un botón para poder descargarlo.

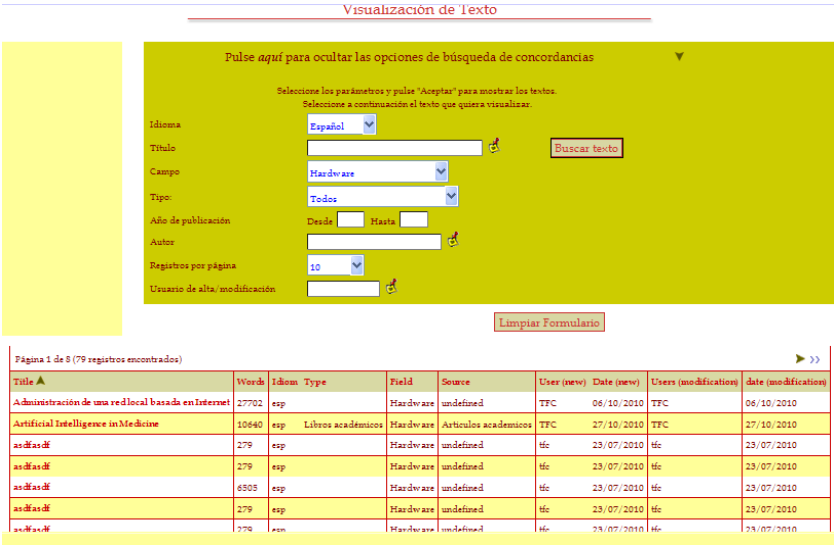

Figura 1. Vista de elección de textos

El **listado** mostrará todas las palabras existentes en los textos seleccionados, junto con el número de veces que aparece en el texto. Cada palabra que aparece en el listado (en azul si está en el glosario), se puede buscar en la RAE[88], en Webopedia[99], en Wikipedia[1010], y en EuroWordNet[1111] (EWN). También es posible mostrar todos los párrafos de los textos en los que aparece esa palabra (es decir, mostrarla en un contexto más amplio), añadirla al glosario o modificar su información en el glosario, como se puede ver en la Figura 2

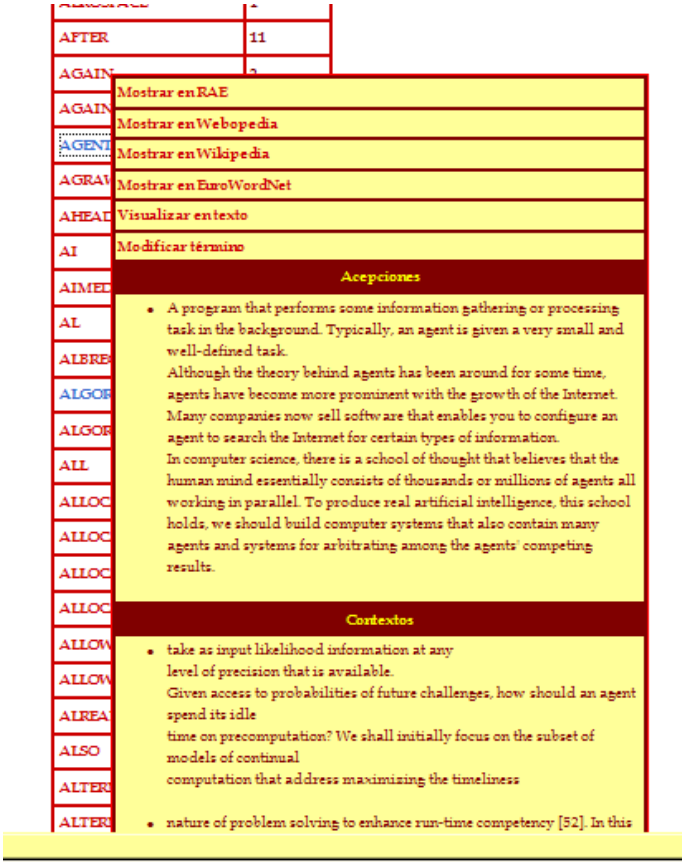

Figura 2. Vista del menú para un término del listado

Las **concordancias** permiten ver las ocurrencias del término buscado en los textos seleccionados, junto con el número de palabras

que se quiere que antecedan o sigan al término en cuestión a la derecha, a la izquierda o a ambos lados (Véase la Figura 3). Esta prestación es similar a la de los concordancieros habituales, aunque incorpora algunas características interesantes. Por ejemplo, si se pincha sobre cualquiera de los términos resaltados en dorado, se muestra en una ventana nueva el párrafo en el que aparece el término. Los resultados (todos o sólo los seleccionados) se pueden guardar en formato xls.

Figura 3. Resultados de la búsqueda de concordancias para “agent”

Además de mostrar las apariciones en contexto de una palabra, Calíope incorpora también la posibilidad de buscar coocurrencias. Esto se puede hacer por término o por categoría gramatical. En el primer caso, por cada término introducido hay que especificar la distancia (exacta o no) al término anterior, junto con el número de palabras que se quiere que antecedan o sigan al término en cuestión a la derecha, a la izquierda o a ambos lados. Los resultados se muestran de manera similar a las concordancias (Véase la Figura 4) y se permiten también las mismas operaciones.

Figura 4. Resultados de búsquedas de coocurrencias para “click on”

En el segundo caso, en lugar de introducir otra palabra, se especifica la categoría gramatical del término que coocurre con el primero. Dado que el corpus está sin anotar, la categoría gramatical se conoce gracias a la integración de EWN en Calíope. En los resultados, además del primer término, se muestran sobre fondo dorado las palabras del entorno que pertenecen a la categoría gramatical requerida. Se pueden hacer las mismas operaciones con ellos que en el caso de las concordancias.

Además de introducir, modificar y eliminar textos, la **administración del corpus** permite gestionar los tipos, campos y fuentes de los textos.

2.2 Operaciones sobre el glosario

Para **consultar** en el glosario, los términos se pueden buscar por inicial, por término (exacto o utilizando caracteres comodín) y por categoría gramatical. Para cada término se indica su lengua y si se trata de un término simple o compuesto. En este apartado es en donde se han incorporado más novedades respecto de una primera versión inicial. Por un lado se han incluido botones con acceso a la RAE, Webopedia, Wikipedia, y EWN. Por otro lado, se cuenta con información lingüística por cada acepción introducida: categoría gramatical; traducción, con el enlace correspondiente si se encuentra el término equivalente en el glosario; definición; contextos de aparición del término en los textos del corpus para esa acepción; y relaciones (tanto jerárquicas como *ad-hoc*) con otros términos. Por cada relación con otro término se muestra el tipo de relación y las observaciones pertinentes, por ejemplo, para ilustrar su sentido. Se puede pinchar sobre el otro término de la relación y ver toda su información almacenada en el glosario en una nueva ventana. Finalmente, también se dispone de la posibilidad de ver una imagen correspondiente a la acepción, que si no está especificada, se buscará con Google, si bien esta búsqueda no garantiza encontrar la imagen adecuada. La Figura 5 muestra una entrada del glosario.

La **administración del glosario** sirve para añadir, modificar y eliminar términos y sus datos. También incluye la administración de las relaciones entre los términos y los tipos de relaciones. Esta última

característica de la aplicación posibilita que el usuario pueda utilizar los tipos de relaciones que más se adecúen a sus necesidades (relaciones *ad-hoc*), en caso de querer emplear tipos de relaciones que no estén preestablecidas, por ejemplo, para reflejar una teoría lingüística concreta (Véase Figura 6).

Figura 5. Información sobre "agent" en el glosario

Figura 6. Vista de creación de relaciones entre términos

2.3 Operaciones sobre el histórico y los usuarios

En lo que respecta a la **administración** de los **usuarios**, las operaciones que se pueden realizar son las habituales de cualquier aplicación web: un usuario normal puede cambiar su contraseña y el privilegiado podrá dar de alta, modificar o eliminar cualquier usuario, asignándole los privilegios adecuados.

Gracias a la **administración** del **histórico** se pueden ver todas las operaciones que se han producido en el sistema, así como el usuario que ha realizado cada una de ellas. Se pueden hacer búsquedas según la acción, el usuario, la entidad sobre la que se ha realizado la acción y la fecha en que se realizó. Por último, se pueden borrar los resultados que se desee.

3. CONCLUSIONES Y LÍNEAS FUTURAS

La herramienta que se ha presentado aquí posee una serie de características que demuestran la idoneidad de Calíope para los fines perseguidos. Entre ellos cabe mencionar los que se exponen a continuación.. Los ejemplos de uso de los términos del glosario se pueden incorporar directamente del corpus a partir de sus concordancias en los textos y eligiendo las adecuadas a la acepción que se esté tratando. Otro rasgo relevante, y novedoso en buena medida es que permite reflejar en el glosario el resultado de algunos análisis del corpus, tarea que no suele ser habitual en los diccionarios en línea ni en los gestores de corpus. Esto se consigue estableciendo distintas relaciones entre los términos, por ejemplo, para indicar las colocaciones que se han logrado identificar tras el estudio de un término. Existe un conjunto predefinido de relaciones que el administrador puede ampliar en caso de ser necesaria una nueva categoría. Respecto a los concordancieros tradicionales, destaca la posibilidad de buscar coocurrencias entre varias palabras, que pueden estar separadas por una distancia máxima dada, y entre una palabra y todas las pertenecientes a una categoría gramatical, tarea que generalmente no está incluida en los concordancieros convencionales.

En resumen, la principal ventaja de Calíope es que todos sus elementos están interconectados, el corpus con el glosario de términos

y estos entre sí, y todo esto se presenta de manera sencilla y comprensible al usuario, sin olvidar el carácter de diccionario visual que le proporciona la inclusión de imágenes.

Por otra parte, es importante destacar que se ha podido comprobar la versatilidad de Calíope como herramienta de ayuda en varios proyectos para las que no estaba inicialmente destinado, como la traducción de textos informáticos y la detección de sentimientos en un corpus de redes sociales.

Sin embargo, cabe mencionar que existen varios aspectos de Calíope en los que se puede seguir trabajando, ya sea mejorando las operaciones existentes, ya sea completándola con nuevas operaciones. De entre todas las posibilidades, en la actualidad se están explorando las siguientes líneas: (a) Ampliación a más idiomas, ya que por el momento *Calíope* está preparada sólo para español e inglés. (b) Incorporación de la anotación proporcionada por Freeling (Padró y Stanilovsky, 2012), dado que la aplicación se beneficiaría ampliamente de la posibilidad de trabajar con un corpus anotado, por ejemplo para mejorar la funcionalidad de coocurrencias y facilitar la detección de términos. (c) Capacidad de tratar el género de sustantivos y adjetivos, y las conjugaciones de los verbos, pues, por ahora, el sistema únicamente distingue el número de los sustantivos y adjetivos. (d) Compatibilidad con múltiples formatos, dado que la aplicación actual sólo maneja los formatos txt y HTML.

NOTAS

[1] http://www.iula.upf.edu/corpus/corpuses.htm
[2] http://bwananet.iula.upf.edu/
[3] http://view.byu.edu/
[4] http://www.lexically.net/wordsmith/
[5] http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/
[6] http://www.termiumplus.gc.ca/
[7] http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl
[8] http://buscon.rae.es/draeI/
[9] http://www.webopedia.com/
[10] http://es.wikipedia.org/wiki/Wikipedia:Portada
[11] http://www.illc.uva.nl/EuroWordNet/

REFERENCIAS BIBLIOGRÁFICAS

Aguado de Cea, G. y Bernardos, Mª S. 2007. “Calíope: herramienta para gestionar un corpus y un glosario de términos informáticos”. *Proceedings of the 6th Annual Conference of the European Association of Languages for Specific Purposes* (AELFE 2007). Lisbon (Portugal).

De Schryver, G.-M. y G. De Pauw. 2007. “Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex”. *Lexikos* 17: 226-246.

Granger, S. 2011. “Learner Corpora”. In: Chapelle C.A., *Language and Technology. The Encyclopedia of Applied Linguistics*, Oxford : Blackwell-Wiley

L’Homme, M.C. 2009. *Manuel DiCoInfo*. [Documento de Internet disponible en http://olst.ling.umontreal.ca/dicoinfo/manuel-DiCoInfo.pdf]

Martín Corral, R. 2012 *Enriquecimiento y mejora de Calíope (V2)* Proyecto Fin de Carrera. Facultad de Informática, Universidad Politécnica de Madrid

Melcuck, I. et al 1984/1988/1992/1999 (4 vols.) *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches lexico-sémantiques I-IV*. Montreal: Les Presses de l'Université de Montréal.

Montaner, A. 2009. Términus. Gestión de corpus y terminología en linea. (Disponible en línea)

Padró, L. y Stanilovsky, E. 2012. “FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA. Estambul (Turquía).

Perales-Escudero, M. y Swales, J. 2011. “Tracing convergence and divergence in pairs of Spanish and English research article abstracts”. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos* (AELFE) 21, 49-7.

Sánchez, A. 2000. “Language Teaching before and after ‘Digitalized Corpora’. Three main issues”. En Cantos, P. & Sánchez, A. *Corpus-based research in English Languaage and Linguistics*, Vol. 9. 1. Murcia: Cuadernos de Filología Inglesa

Sinclair, J.M. (Ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London/Glasgow: Collins ELT.

Vossen, P. 1998. *EuroWordNet*. Final Report. http://www.lsi.upc.edu/~escudero/wsd/98-ewn2728.pdf

Walker, K. 1999. Using Genre Theory to Teach Students Engineering Lab Report Writing: *IEEE Transactions on Professional Communication*, 42, 1, 12-19.